Samuel Ady Sanjaya

Refining Baby Cry Classification using Data Augmentation (Time-Stretching and Pitch-Shifting), MFCC Feature Extraction...

Quick Submit

Quick Submit

Universitas Multimedia Nusantara

Document Details

Submission ID

trn:oid:::1:3308578332

Submission Date

Aug 5, 2025, 3:56 PM GMT+7

Download Date

Aug 5, 2025, 3:59 PM GMT+7

File Name

486_Refining_Baby_Cry_Classification_using_Data_Augmentation.pdf

File Size

3.2 MB

7 Pages

5,613 Words

30,405 Characters

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Exclusions

- 1 Excluded Source

Match Groups

57 Not Cited or Quoted 14%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

11%	Internet sources
10%	Publications
5%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **57** Not Cited or Quoted 14%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

11%	Internet sources
10%	Publications
5%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	eitca.org	1%
2	Internet	ijcs.stmikindonesia.ac.id	1%
3	Internet	www.mdpi.com	<1%
4	Publication	"CONMEDIA 2023 Table of Contents", 2023 7th International Conference on New ...	<1%
5	Publication	Alyaa Hamel Sfayyih, Nasri Sulaiman, Ahmad H. Sabry. "Non-invasive diagnosis of...	<1%
6	Publication	Aburakhia, Sulaiman A.S.. "Data-Driven Vibration-Based Condition Monitoring : F...	<1%
7	Student papers	University of Warwick	<1%
8	Publication	Nidhi Chakravarty, Mohit Dua. "An improved feature extraction for Hindi languag...	<1%
9	Student papers	University of Northampton	<1%
10	Publication	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ...	<1%

11	Student papers	Amrita Vishwa Vidyapeetham	<1%
12	Internet	monarch.qucosa.de	<1%
13	Internet	researchrepository.universityofgalway.ie	<1%
14	Internet	www.econstor.eu	<1%
15	Publication	Chappidi Suneetha, Raju Anitha. "Speech based emotion recognition by using a fa...	<1%
16	Internet	fileadmin.cs.lth.se	<1%
17	Student papers	Liverpool John Moores University	<1%
18	Publication	Md. Abul Ala Walid, Pintu Chandra Shill, SM Tamim Mahmud. "Comparative Analy...	<1%
19	Student papers	University of Stellenbosch, South Africa	<1%
20	Student papers	University of Southampton	<1%
21	Internet	kluedo.ub.uni-kl.de	<1%
22	Publication	Shamshair Ali, Rubina Ghazal, Nauman Qadeer, Oumaima Saidani et al. "A novel ...	<1%
23	Student papers	University of Central Florida	<1%
24	Internet	di.univ-blida.dz	<1%

25	Internet	icbsii.in	<1%
26	Internet	network.bepress.com	<1%
27	Internet	reunir.unir.net	<1%
28	Publication	Aleksandr Kulikov, Pavel Pelevin, Anton Loskutov. "Validation of a simulation mo...	<1%
29	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artific...	<1%
30	Publication	Arya Shah, Amanpreet Singh, Artika Singh. "Chapter 6 Audio Classification of Eme...	<1%
31	Internet	jmcms.s3.amazonaws.com	<1%
32	Internet	jurnal.polibatam.ac.id	<1%
33	Internet	www.lmf.bgu.tum.de	<1%
34	Publication	"Proceedings of 4th International Conference on Recent Trends in Machine Learn...	<1%
35	Publication	Edraki, Amin. "Prediction and Enhancement of Speech Intelligibility in Challengin...	<1%
36	Publication	Shreya Shreya, Raghav Goel, Priyanshi Katariya, Harshit Kamboj, Nirbhay Kashya...	<1%
37	Publication	Tayyip Ozcan, Hafize Gungor. "Baby Cry Classification Using Structure-Tuned Artif...	<1%
38	Publication	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challeng...	<1%

39	Internet	dspace.vut.cz	<1%
40	Internet	eprints.qut.edu.au	<1%
41	Internet	estudogeral.sib.uc.pt	<1%
42	Internet	gyan.iitg.ac.in	<1%
43	Internet	isassymposium.org	<1%
44	Internet	www.igi-global.com	<1%
45	Internet	www.isca-archive.org	<1%
46	Internet	www.researchgate.net	<1%

Refining Baby Cry Classification using Data Augmentation (Time-Stretching and Pitch-Shifting), MFCC Feature Extraction, and LSTM Modeling

1st Vinka Bella
Faculty of Technology and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
vinka.bella@student.umn.ac.id

2nd Samuel Ady Sanjaya
Faculty of Technology and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
samuel.ady@umn.ac.id

***Abstract*— Babies, in their early developmental stages, are unable to communicate their needs through language. When they seek to convey discomfort or express their needs, they primarily resort to crying. The dataset in this research contains 497 sounds of baby cries that are divided into 5 categories: hungry, belly pain, discomfort, burping, and tired. In this context, this paper proposes an innovative approach to address this challenge by employing Long Short-Term Memory (LSTM) networks in combination with Mel-Frequency Cepstral Coefficients (MFCC) as feature extraction for the baby cries analysis. The study focuses on developing a model capable of discerning the underlying messages within a baby's cry by leveraging the acoustic characteristics of the cry and the temporal dependencies inherent in the data. Apart from combining MFCC and LSTM, we also add two data augmentations specifically time stretching and pitch shifting to improve model performance. After the data has been successfully augmented, we extract MFCC features from recorded cry samples, and these features are used as input data for the LSTM neural network. The LSTM Model is created with 12 hidden layers and 30 epochs to run. The combination of data augmentation using time-stretch and pitch-shifting resulted in 96% accuracy on validation results, compared to only 72% accuracy on non-augmented data. The model with augmented data also has resulted in better loss which indicates the model is trained better using from larger dataset. In conclusion, it can be said that the combination of data augmentation and feature extraction has a significant impact on a model's ability to learn.**

Keywords—baby cry, MFCC, LSTM, pitch shifting, time stretching.

I. Introduction

The journey of parenthood is a profound and transformative experience, especially for new parents who are embarking on the voyage of raising their first child. As they embark on this life-altering adventure, the joy and wonder of parenthood are often intermingled with moments of uncertainty and trepidation [1]. A recurring and a challenge faced by new parents, one that transcends cultural and geographical boundaries, is deciphering the unspoken language of their infants. Babies, in the early stages of their development, lack the capacity to articulate their desires and requirements through conventional speech. Instead, they employ a primal and instinctive form of communication—crying [2]. The cry of a baby becomes the medium through which they convey their feelings, needs, and discomfort. It is the universal language of infancy, a non-verbal expression that calls upon the caregiver's attention and nurturance [3].

In this complex baby cry language, parents find it difficult to understand the differences between a hungry cry, a cry of pain, a cry of disturbance, a cry signaling the need for burping, and many more expressions of emotions that, even though they don't use words, are full of meaning [4]. Each cry serves as a subtle indicator, a plea from the infant for something vital—nourishment, comfort, solace, or relief from distress. In response, parents strive to decode these vocal cues, aspiring to provide their child with the best possible care and support [5]. This paper delves into the challenges confronted by new parents in interpreting the multifaceted language of baby cries, exploring the fundamental needs and emotions that underlie these vocal expressions. It also introduces an innovative approach aimed at harnessing technology and advanced methods to aid parents in comprehending and responding to the intricate symphony of their baby's cries. Through this, we seek to enhance the caregiving experience, fostering stronger bonds and improved well-being for both parents and their precious infants.

In the quest to decipher the enigmatic language of baby cries, previous generations of parents have relied on a combination of psychological intuition and sociological understanding, passed down through cultural wisdom and familial traditions. However, in this modern era, we find ourselves at the cusp of a transformative evolution. Advancements in technology have introduced innovative tools and approaches, which promise to aid and empower parents in unprecedented ways [6]. In the age of Artificial Intelligence (AI), a new frontier emerges where AI systems stand ready to assist and guide parents in understanding their babies' needs with remarkable precision [7]. Indeed, technological strides in sound classification and machine learning have opened doors to an exciting realm of possibilities [8][9]. Previous research endeavors have explored sound classification, including the challenging domain of baby cry sound classification. These investigations have laid the foundation for the current study, offering valuable insights and methodologies upon which we build our framework [10].

In this paper, we present a comprehensive framework for baby cry sound classification, combining state-of-the-art techniques to discern and respond to an infant's emotional and physical requirements. Our proposed methodology encompasses a multi-step process using data augmentation, feature extraction, and deep learning modeling. Recognizing the diversity and subtlety of baby cries, we employ data augmentation techniques such as pitch shifting and time stretching. These methods enrich our dataset, capturing the broad spectrum of cry variations, and ensuring a robust model capable of accommodating the uniqueness of each infant's cry. MFCC Feature Extraction: We leverage Mel-Frequency Cepstral Coefficients (MFCC) to extract discriminative features from the augmented cry data [11]. These features encapsulate the essential spectral information inherent in the cries, enabling the subsequent modeling phase to be both effective and efficient [12]. Modeling with LSTM: To decode the complex patterns within baby cries, we employ Long Short-Term Memory (LSTM) neural networks. LSTMs are well-suited for capturing temporal dependencies, making them ideal for understanding the subtle changes in pitch, rhythm, and duration within the cries [13]. This model is at the heart of our framework, serving as the cornerstone for classification.

Through this combination of modern technology and traditional caregiving, our model endeavors to help parents to understand their babies' needs. As we embark on this exploration of AI-assisted baby cry analysis, we anticipate a transformative shift in parenting dynamics, fostering enhanced caregiving and, ultimately, a more profound connection between parents and their infants.

II. Methods

Our methodology involves data collection, data augmentation, data visualization through wave plots and spectrograms, the application of Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction, and the implementation of LSTM modeling and evaluation. These components collectively form the foundation of our approach to understanding and classifying baby cries, promising to enhance the caregiving experience for parents and caregivers.

A. *Data Collection*

Data collection for this study was a meticulous process designed to ensure the quality and relevance of the dataset. The following steps detail our data collection methodology: Source Selection: The initial source of our data was the "Donate-a-Cry" repository available on GitHub, an open-source platform [14]. This repository was chosen for its comprehensive collection of baby cry recordings, because the dataset is already cleaned with these following step:

1) *Data Format Standardization:* To establish uniformity within the dataset, all collected audio files were converted to WAV format. This conversion maintained consistent bit and sampling rates of 128 kbps and 8kHz, respectively. This standardization was crucial for ensuring that all audio data would be compatible for subsequent analysis.
2) *Data Label Refinement*: To align the dataset with the Dunstan Baby Language (DBL) categories [15], data tagged with descriptors such as "lonely," "scared," and "unknown" were removed. Additionally, data tagged with "cold" and "hot" were merged into a single category labeled "discomfort." This classification process enabled us to categorize cries according to fundamental emotional states and needs.
3) *Manual Noise Elimination:* To ensure the dataset's purity, non-cries, or sounds that did not represent genuine baby cries, were meticulously identified and removed manually. This process involved listening to each audio sample and eliminating any non-relevant data, such as white noise, instances of baby chat, or adults mimicking baby cries.

By following this systematic data collection approach, we aimed to construct a reliable and well-curated dataset that would serve as the foundation for our subsequent analysis and classification of baby cries. This rigorous methodology ensured that the dataset accurately represented the cries of infants in various emotional states and needs, laying the groundwork for our research endeavors.

B. *Data Augmentation*

Data augmentation is a crucial step in the data preprocessing pipeline, especially when working with limited datasets [16]. In this study, the collected dataset, while valuable, did not provide enough samples for each cry category. To address this limitation and to ensure the dataset's comprehensiveness, we employed data augmentation techniques, specifically time stretching and pitch shifting.

1) *Time Stretching*: Time stretching is a technique used to alter the duration of an audio signal while maintaining its pitch [17]. It is particularly useful when the available data is insufficient, as it allows for the creation of variations of the existing audio clips without changing their fundamental characteristics [18]. The time-stretching process involves resampling the audio signal to either a higher or lower sampling rate, effectively modifying the playback speed. To stretch the time of an audio signal by a factor of 'α,' show on equation (1).

   New sample n in the stretched signal:

   $$x\_new(n) = x(\alpha n) \tag{1}$$

   Where,

   $x(n)$ is the original audio signal.
   $x\_new(n)$ is the stretched signal.
   α is the stretching factor ($\alpha > 1$ for stretching, $\alpha < 1$ for compression).

   By altering the value of α, we extend or compress the duration of the audio signal as needed. In our case, time stretching allowed us to create additional cry samples for each category, thereby enriching our dataset without affecting the pitch of the cries.
2) *Pitch Shifting*: Pitch shifting, on the other hand, focuses on altering the pitch or frequency of an audio signal while keeping its duration constant [19], [20]. This technique is valuable for generating variations in the pitch of cries, which can be informative in understanding the nuances of infant vocalizations [21]. The pitch-shifting process can be described on the equation (2).

   New sample n in the shifted signal:

   $$x\_new(n) = x(n/\beta) \tag{2}$$

   Where,

   $x(n)$ is the original audio signal.
   $x\_new(n)$ is the shifted signal.
   β is the pitch shifting factor ($\beta > 1$ for pitch increase, $\beta < 1$ for pitch decrease).

Adjusting the value of β allows to raise or lower the pitch of the audio while preserving its original timing. In our research, pitch shifting was applied to create variations in pitch within the cry dataset, enhancing its diversity and capturing a broader spectrum of infant vocal expressions.

By combining time stretching and pitch shifting techniques, we successfully augmented our dataset, addressing the limitations in the quantity of data for each cry category. This augmentation not only increased the dataset's size but also introduced variations that can be essential for robust model training and improved classification accuracy.

C. Data Visualization: Waveplot and Spectrogram

Data visualization plays a pivotal role in the analysis and understanding of baby cry data. In our study, we utilized two key visualization techniques: waveplots and spectrograms, to gain insights into the temporal and spectral characteristics of infant cries [22].

1) *Waveplots*: A waveplot is a time-domain representation of an audio signal, which provides a visual representation of the signal's amplitude as a function of time. In the context of baby cry analysis, waveplots are valuable for understanding the temporal dynamics of the cries. A waveplot displays the original audio waveform in a two-dimensional graph, with time on the x-axis and amplitude on the y-axis. Peaks in the waveplot correspond to instances of high amplitude, indicating louder portions of the cry, while valleys represent softer or quieter segments.

2) *Spectrograms*: While waveplots provide insight into the time domain, spectrograms offer a window into the frequency domain of an audio signal. A spectrogram is a visual representation that illustrates how the signal's frequency content changes over time [23]. It breaks down the audio signal into its constituent frequencies and shows how their amplitudes vary with time.

In our research, the combination of waveplots and spectrograms served as a powerful tool for visualizing and understanding the temporal and spectral aspects of baby cries. This comprehensive approach provided the foundation for subsequent data preprocessing and analysis, enabling us to extract meaningful features for classification and interpretation.

D. Feature Extraction using MFCC

In the field of sound classification, feature extraction is a crucial step in the data preprocessing pipeline, aimed at transforming raw audio signals into a format that machine learning algorithms can effectively work with. One of the most widely used and effective techniques for this purpose is Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are a representation of the short-term power spectrum of an audio signal that mimics the human auditory system's sensitivity to sound [24]. The MFCC extraction process can be broken down into several steps:

1) *Frame Segmentation*: The audio signal is divided into short overlapping frames, typically ranging from 20 to 40 milliseconds in duration. These frames capture the non-stationary characteristics of sound over time.

2) *Windowing*: Each frame is multiplied by a windowing function (e.g., Hamming window) to reduce spectral leakage. This ensures that the data at the edges of the frames does not have as much influence on the analysis. The equation is shown in equation (3).

$$w_n = 0.5\left(1-\cos\frac{2\pi n}{N-1}\right), 0 \le \mathrm{n} \le \mathrm{N}-1 \quad (3)$$

3) *Fast Fourier Transform (FFT)*: A Fourier Transform is applied to each frame to convert it from the time domain to the frequency domain. This results in a power spectrum that represents the distribution of power across different frequency bands [25].

4) *Mel Filterbank*: A set of triangular filters, arranged on a Mel scale, is applied to the power spectrum. The Mel scale approximates the human auditory system's response to different frequencies. This step transforms the power spectrum into a set of filterbank energies, which represent how much energy is present in different frequency bands. The Mel Filterbank equation is shown in equation (4).

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \quad (4)$$

5) *Logarithm*: The logarithm is applied to the filterbank energies. This is done to mimic the non-linear response of the human ear to loudness. Taking the logarithm also helps to reduce the dimensionality of the feature vectors.

6) *Discrete Cosine Transform (DCT)*: The DCT is applied to the logarithm of the filterbank energies. This decorrelates the coefficients, resulting in the final set of MFCCs, which are used as feature vectors for sound classification.

In the context of baby cry analysis, MFCCs offer a compact yet highly informative representation of cry sounds, capturing the essential spectral characteristics that distinguish different emotional states and needs. These coefficients serve as the foundation for subsequent classification algorithms, helping to identify and interpret the underlying meanings within infant vocalizations [26].

E. LSTM Modeling and Evaluation

LSTM is a specialized neural network architecture designed to handle sequences of data with long-term dependencies. LSTM accomplishes this by introducing a more complex memory cell that can selectively retain and forget information, making it well-suited for processing sequential data such as audio signals [27]. In the context of baby cry classification, LSTM networks are typically designed as a combination of LSTM layers and dense layers. The LSTM layers are responsible for capturing the sequential information in the cry signals, while the dense layers at the output end transform the learned features into classification decisions [28]. The model is trained using a labeled dataset, where each cry is associated with a specific category (e.g., hungry, discomfort, pain).

The Long Short-Term Memory (LSTM) unit is a complex recurrent neural network (RNN) architecture that is specifically designed to capture long-range dependencies in sequential data. The equations governing an LSTM unit are shown in equation (5)-(10):

Input Gate (i_t):

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i \quad (5)$$

Forget Gate (f_t):

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f \quad (6)$$

Output Gate (o_t):

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o \quad (7)$$

New Memory Network ($\widehat{C}_t$):

$$\widehat{C}_t = \tan h\,(W_C \cdot [h_{t-1}, X_t] + b_C) \quad (8)$$

Cell State (C_t):

$$C_t = i_t \cdot \widehat{C}_t + f_t \cdot C_{t-1} \quad (9)$$

Hidden State (h_t):

$$h_t = o_t \cdot \tan h\,(C_t) \quad (10)$$

Where,

i_t, f_t, and o_t are the input, forget, and output gates, respectively.

$\widehat{C}_t$ represents the candidate cell state, which is the new information that can be added to the cell state.

C_t is the cell state.

h_t is the hidden state.

X_t is the input at time step t.

h_{t-1} is the previously hidden state.

W_i, W_f, W_o, W_c, b_i, b_f, b_o and b_c are the weight matrices and bias vectors for the gates and cell state, which are learned during training. The model will be evaluated using accuracy and loss metrics. The loss metrics can also be used to evaluate the model other than the accuracy [29].

III. Result and Discussions

In this section, we delve into the results and discussions stemming from a comprehensive approach to this task, where the primary focus was on harnessing the capabilities of data augmentation, MFCC (Mel-Frequency Cepstral Coefficients) feature extraction, and LSTM (Long Short-Term Memory) modeling.

A. *Data Collection*

In our dataset, we observe a notable class imbalance among the different labels that describe the reasons behind a baby's cry. The most prominent label is 'hungry,' with a count of 382, while other labels such as 'discomfort,' 'tired,' 'belly_pain,' and 'burping' have significantly lower counts of 27, 24, 16, and 8. This class imbalance can pose a significant challenge when training a machine learning model for baby cry classification. To address this class imbalance, we employed a technique known as data augmentation. Data augmentation involves artificially increasing the size of the minority classes by creating additional training examples through various transformations and modifications of the existing data. In the context of baby cry classification, data augmentation allows us to generate more instances of the less frequent labels, thus mitigating the class imbalance issue.

B. *Data Visualization: Waveplot and Spectrogram*

The visualization of a sound waveplot is used to observe the shape of the sound wave of baby cry, which indicates the amplitude of the audio signal presented in the time domain.

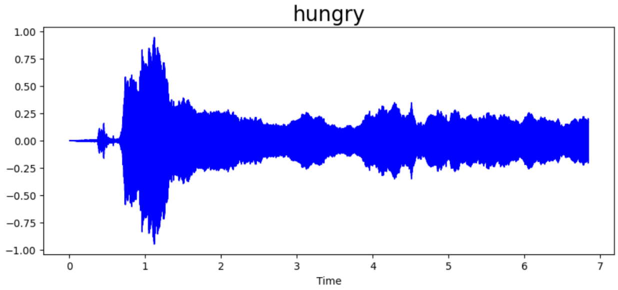


Fig. 1. Waveplot of an original dataset of hungry label

Waveplots are used to identify the fundamental frequency and pitch of a sound as shown in Fig 2. They can be helpful in detecting distortion or noise in the sound, thus enabling filtering. Visualization using a spectrogram is done to understand the sound patterns produced by a baby as shown in Fig 3. The image shows the distibution of frequency (Hz) and the decibel (dB) over time.

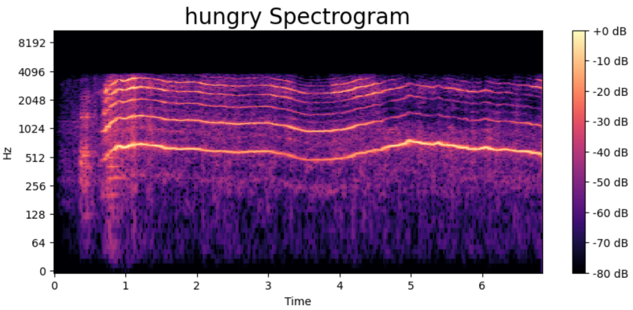


Fig. 2. Spectogram of original dataset of hungry label

C. *Data Augmentation*

Fig 3 from Data Visualization is the original data, while Fig 4-7 illustrates a crucial aspect of our study's data preprocessing techniques, showcasing the necessity of data augmentation to address the challenges posed by imbalanced data. To mitigate this imbalance and ensure that the model can effectively learn from and classify all classes, we employed data augmentation strategies. The time stretching and pitch shifting techniques are vital for generating additional instances of minority classes, enabling the model to build a more balanced and robust understanding of the different baby cry categories.

1) *Time Stretching:* One of the fundamental data augmentation techniques we applied in our study is time stretching. Time stretching, as demonstrated in Fig 4 dan Fig 5, involves the modification of audio sequences to either extend or compress their duration. This process preserves the original pitch and spectral characteristics while altering the temporal structure.

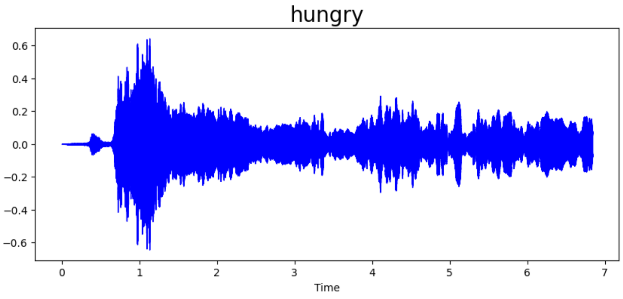


Fig. 3. Waveplot of Time-Stretching

By comparing the modified 'hungry' class to the original data, we can observe the time-stretched waveform, which expands the sound sequence. The comparison reveals the advantage of time stretching in creating variations within the 'hungry' class, enabling the model to better capture the temporal nuances of this type of cry.

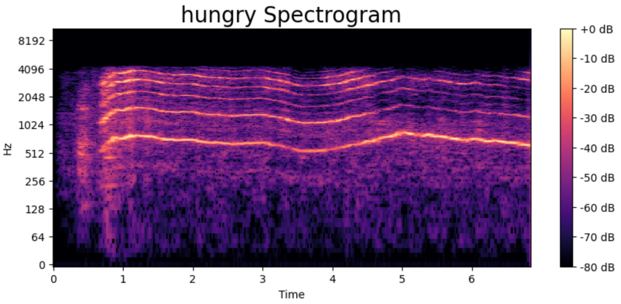


Fig. 4. Spectogram of Time-Stretchhing

2) *Pitch Shifting:* In addition to time stretching, we employed pitch shifting as another data augmentation technique. Pitch shifting, as depicted in Fig 6 and Fig 7, focuses on altering the fundamental frequency of an audio signal while preserving its temporal structure. In our comparison with the original data, we can see the effects of pitch shifting on the 'tired' class, which results in a shift in the pitch of the cry.

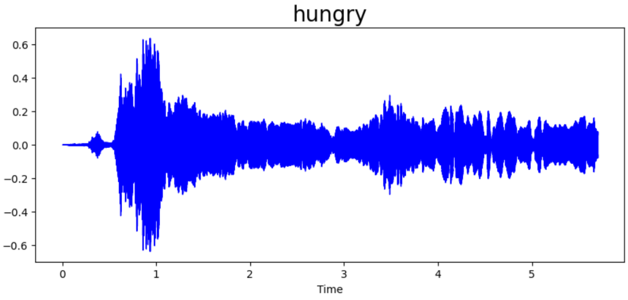


Fig. 5. Waveplot of Pitch Shifting

This augmentation technique introduces variations in the pitch domain, helping the model learn to recognize 'hungry' cries with different tonal characteristics. These variations are essential for enhancing the model's ability to distinguish between different baby cry categories, ultimately improving its overall classification performance.

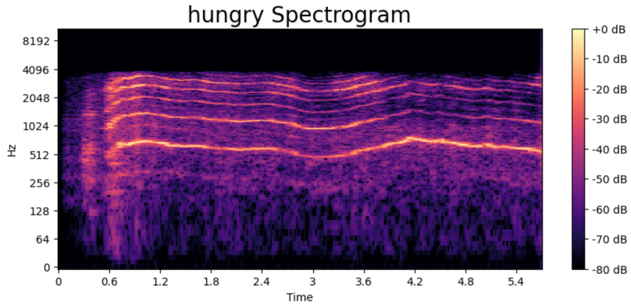
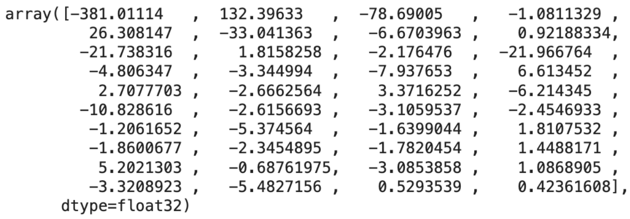


Fig. 6. Spectogram of Pitch Shifting

After the augmentation, the data is tripled from 500 data to 1500 data, and this augmented data will be used to model the sound classification.

D. Feature Extraction using MFCC

Feature extraction from a baby crying dataset is performed as signal processing, before applying the LSTM algorithm for sound classification. By adopting the Mel-frequency cepstral coefficients (MFCC) method, the audio is transformed into an array form that allows for a more machine-friendly representation in recognizing characteristics in the voice signal, such as melody, rhythm, and special pattern.



```
array([-381.01114   ,  132.39633   ,  -78.69005   ,   -1.0811329 ,
         26.308147  ,  -33.041363  ,   -6.6703963 ,    0.92188334,
        -21.738316  ,    1.8158258 ,   -2.176476  ,  -21.966764  ,
         -4.806347  ,   -3.344994  ,   -7.937653  ,    6.613452  ,
          2.7077703 ,   -2.6662564 ,    3.3716252 ,   -6.214345  ,
        -10.828616  ,   -2.6156693 ,   -3.1059537 ,   -2.4546933 ,
         -1.2061652 ,   -5.374564  ,   -1.6399044 ,    1.8107532 ,
         -1.8600677 ,   -2.3454895 ,   -1.7820454 ,    1.4488171 ,
          5.2021303 ,   -0.68761975,   -3.0853858 ,    1.0868905 ,
         -3.3208923 ,   -5.4827156 ,    0.5293539 ,    0.42361608],
      dtype=float32)
```

Fig. 7. Applied MFCC from Sound-Wave Dataset

The soundwave transformation to array shown in Fig 8, the results of feature extraction with MFCC. Where from the wav file is generated into numbers of array.

E. LSTM Modeling and Evaluation

The model in this research is a sophisticated deep learning architecture for the analysis of sequential data. This model is characterized by its depth and the increased number of parameters to learn. The design includes a 1D convolutional layer with 64 filters and a 3-element kernel, which, in combination with max-pooling, allows the model to extract intricate features from the input data.

However, the distinguishing feature of the model is its recurrent layers, which consist of three LSTM (Long Short-Term Memory) layers. These LSTM layers have progressively larger numbers of units, which allows the model to learn and capture increasingly intricate temporal dependencies in the data. By setting the return sequences parameter to True for the first two LSTM layers, the architecture retains sequential output data, preserving essential temporal information. The detail of the model is shown in Table 1.

Table 1. LSTM Model Built Layers

Layer	Layers		
	Layer (type)	Output Shape	Params#
1	Convolutional 1D	38, 64	256
2	MaxPoolingID	19, 64	0
3	LSTM I	19, 256	328,704
4	LSTM 2	19, 128	197,120
5	LSTM 3	64	49,408
6	Dense III	256	16,640
7	DroupOut III	256	0
8	Dense IV	128	32,896
9	Dropout IV	128	0
10	Dense V	64	8,256
11	Dropout V	64	0
12	Dense VI	5	325

Following the LSTM layers, dense layers with ReLU activation functions are applied. Remarkably, dropout layers, with a relatively high rate of 0.4, are introduced between the dense layers. These dropout layers serve the purpose of regularization by randomly deactivating 40% of the neurons during training, thus preventing overfitting and enhancing the model's generalization capabilities. Finally, the model concludes with a dense layer featuring five output units, equipped with a softmax activation function, indicative of a multi-class classification task. This intricate architecture is designed to delve into complex, multi-layered temporal patterns within the data. Accuracy in training and validation can have different optimal parameters. This was caused by the nature of neural network training and the trade-off among fitting the training data and generalizing to new data or validation. During training, the model learns from the training data.

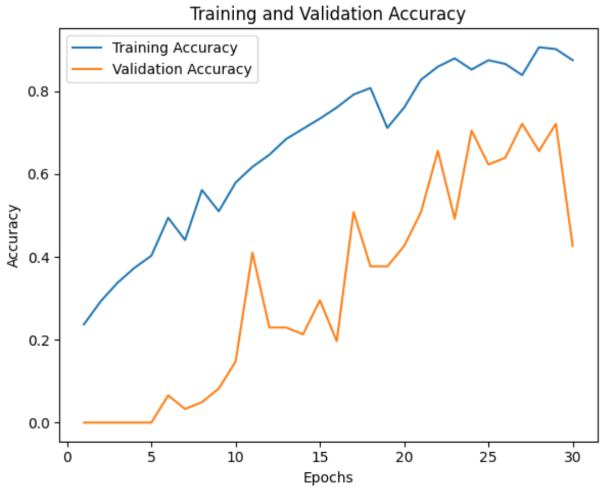


Fig. 8. Training and Validation Accuracy non-Augmented Data

In our initial experiment, we embarked on a modeling endeavor without employing any data augmentation techniques, utilizing a dataset comprising 500 samples. The outcomes of this experiment are thoughtfully illustrated in Fig 9 and Fig 10, offering a visual representation of the model's performance metrics. As we observe in these figures, the accuracy curve attains its zenith at a commendable 72%, while the loss curve reaches its nadir at 1.04, both occurring at the 30th epoch. This data unequivocally underscores the potential limitations of utilizing non-augmented data, revealing a clear plateau in the model's performance without the benefit of data augmentation techniques.

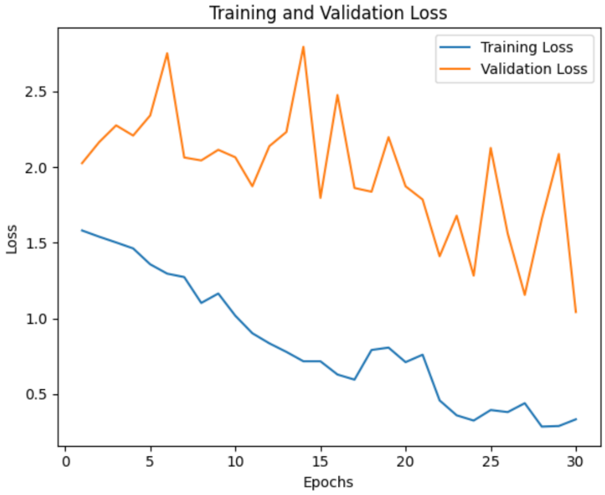


Fig. 9. Training and Validation Loss non-Augmented Data

In our second experiment, a pivotal shift occurred as we introduced the transformative technique of data augmentation, expanding our dataset to a robust 1500 samples. Fig 11 and Fig 12 unveil the compelling transformation in the model's performance brought about by this augmentation. Notably, the accuracy surged to an impressive 96%, representing a substantial leap from the previous experiment, and the loss plummeted to a remarkable 0.35, showcasing the model's enhanced efficiency and precision. These results underscore the significant impact of data augmentation in bolstering both accuracy and efficiency, highlighting its pivotal role in optimizing the model's learning process and predictive capabilities.

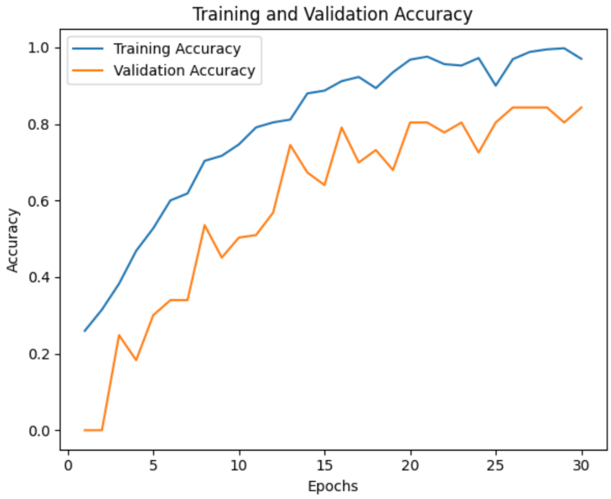


Fig. 10. Training and Validation Accuracy Augmented Data

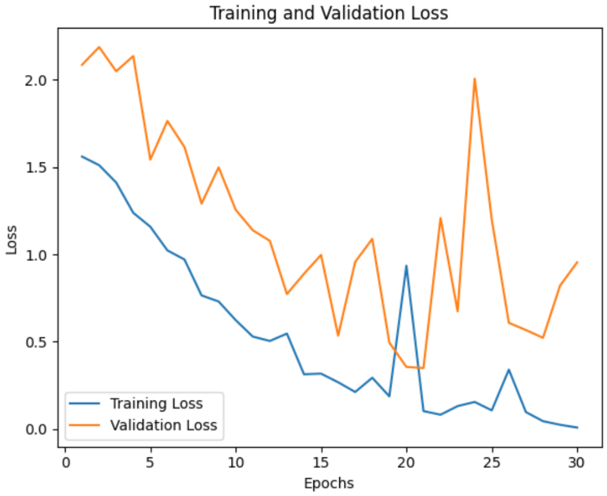


Fig. 11. Training and Validation Loss Augmented Data

By leveraging accelerated training iterations, models that incorporate augmented data can achieve markedly enhanced levels of accuracy. This accelerated training process empowers the models to glean more comprehensive insights from the data, resulting in substantially improved predictive performance and precision. In essence, faster iterations enable these models to learn more efficiently and effectively, ultimately leading to superior outcomes in terms of their accuracy.

IV. Conclusions

In conclusion, the creation of the baby cry classification model involved the augmentation of data, the extraction of features through the use of MFCC, and extensive LSTM modeling. The LSTM model was constructed using 12 hidden layers and more than 50 training epochs with the Adam optimizer. The final model performed admirably, with an accuracy of 96% which is much better than non-augmented data only resulting in 72%. Even though these results are encouraging, more research will lead to advancements in the future. More specifically, if the dataset is larger and more comprehensive, the model will be able to learn more. Examining various feature extraction techniques may also reveal the greatest methods for enhancing the model and create new avenues for study and advancement. This work lays the foundation for future advancements in the field of baby cry categorization, with the possibility for even more dependable and accurate models to be developed.

ACKNOWLEDGMENT

This research received support from the Institution of Research and Community Services at Universitas Multimedia Nusantara. We extend our appreciation to our colleagues at the Big Data Laboratory within the Information Systems Department at Universitas Multimedia Nusantara, whose valuable input and expertise greatly enriched the research.

REFERENCES

[1] S. P. Dewi, A. L. Prasasti, and B. Irawan, "Analysis of LFCC Feature Extraction in Baby Crying Classification using KNN," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, Nov. 2019, pp. 86–91. doi: 10.1109/IoTaIS47347.2019.8980389.

[2] D. Widhyanti and D. Juniati, "Classification of Baby Cry Sound Using Higuchi's Fractal Dimension with K-Nearest Neighbor and Support Vector Machine," *J. Phys. Conf. Ser.*, vol. 1747, no. 1, p. 012014, Feb. 2021, doi: 10.1088/1742-6596/1747/1/012014.

[3] A. A. Ikhsania, "10 Penyebab Bayi Menangis Terus dan Cara Mengatasinya," *PT. Nutricia Indonesia Sejahtera*, 2023. https://www.nutriclub.co.id/artikel/pola-asuh-anak/bayi/penyebab-dan-cara-mengatasi-bayi-menangis (accessed Oct. 04, 2023).

[4] R. I. TUDUCE, M. S. RUSU, H. CUCU, and C. BURILEANU, "Automated Baby Cry Classification on a Hospital-acquired Baby Cry Database," in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2019, pp. 343–346. doi: 10.1109/TSP.2019.8769075.

[5] L. Novamizanti, A. L. Prasasti, and B. S. Utama, "Study of Linear Discriminant Analysis to Identify Baby Cry Based on DWT and MFCC," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 982, no. 1, p. 012009, Dec. 2020, doi: 10.1088/1757-899X/982/1/012009.

[6] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 9411–9457, Mar. 2021, doi: 10.1007/s11042-020-10073-7.

[7] X. Ren, "Research on a software architecture of speech recognition and detection based on interactive reconstruction model," *Int. J. Speech Technol.*, vol. 24, no. 1, pp. 87–95, Mar. 2021, doi: 10.1007/s10772-020-09770-3.

[8] T. International, "What is audio classification?," *International, TELUS*, 2022. What is audio classification? (accessed Oct. 06, 2023).

[9] D. A. Kristiyanti and M. Wahyudi, "Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Aug. 2017, pp. 1–6. doi: 10.1109/CITSM.2017.8089278.

[10] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.

[11] S. Y. Yusdiantoro and T. B. Sasongko, "Implementasi Algoritma MFCC dan CNN dalam Klasifikasi Makna Tangisan Bayi," *Indones. J. Comput. Sci.*, vol. 12, no. 4, Aug. 2023, doi: 10.33022/ijcs.v12i4.3243.

[12] Y. Yohannes and R. Wijaya, "Klasifikasi Makna Tangisan Bayi Menggunakan CNN Berdasarkan Kombinasi Fitur MFCC dan DWT," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 599–610, Jun. 2021, doi: 10.35957/jatisi.v8i2.470.

[13] ProjectPro, "The ultimate guide to building your own LSTM models," *ProjectPro*, 2023. https://www.projectpro.io/article/lstm-model/832#mcetoc_1gsj4do3ju (accessed Oct. 09, 2023).

[14] A. N. Rizal, J. Jan, L. Dürlich, and S. Chen, "Donate a Cry Corpus," 2019.

[15] M. Lockhart-Bouron *et al.*, "Infant cries convey both stable and dynamic information about age and identity," *Commun. Psychol.*, vol. 1, no. 1, p. 26, Oct. 2023, doi: 10.1038/s44271-023-00022-z.

[16] Y. Jeong, J. Kim, D. Kim, J. Kim, and K. Lee, "Methods for Improving Deep Learning-Based Cardiac Auscultation Accuracy: Data Augmentation and Data Generalization," *Appl. Sci.*, vol. 11, no. 10, p. 4544, May 2021, doi: 10.3390/app11104544.

[17] S. Wei, S. Zou, F. Liao, and W. Lang, "A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification," *J. Phys. Conf. Ser.*, vol. 1453, no. 1, p. 012085, Jan. 2020, doi: 10.1088/1742-6596/1453/1/012085.

[18] N. Akaishi, K. Yatabe, and Y. Oikawa, "Improving Phase-Vocoder-Based Time Stretching by Time-Directional Spectrogram Squeezing," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095348.

[19] W. Zhao and B. Yin, "Environmental sound classification based on pitch shifting," in *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*, Jan. 2022, pp. 275–280. doi: 10.1109/SCSET55041.2022.00070.

[20] X. Yuanchao, C. Zhiming, and K. Xiaopeng, "Improved pitch shifting data augmentation for ship-radiated noise classification," *Appl. Acoust.*, vol. 211, p. 109468, Aug. 2023, doi: 10.1016/j.apacoust.2023.109468.

[21] A. Madhu and S. K., "EnvGAN: a GAN-based augmentation to improve environmental sound classification," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6301–6320, Dec. 2022, doi: 10.1007/s10462-022-10153-0.

[22] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech 2021*, Aug. 2021, pp. 571–575. doi: 10.21437/Interspeech.2021-698.

[23] V. Research, "What is a Spectrogram? - Signal Analysis," *Vibration Research*, 2023. https://vibrationresearch.com/blog/what-is-a-spectrogram/ (accessed Oct. 05, 2023).

[24] L. S. Foo, W.-S. Yap, Y. C. Hum, Z. Kadim, H. W. Hon, and Y. Kai Tee, "Real-Time Baby Crying Detection in the Noisy Everyday Environment," in *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, Aug. 2020, pp. 26–31. doi: 10.1109/ICSGRC49013.2020.9232488.

[25] M. S. Imran, A. F. Rahman, S. Tanvir, H. H. Kadir, J. Iqbal, and M. Mostakim, "An Analysis of Audio Classification Techniques using Deep Learning Architectures," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Jan. 2021, pp. 805–812. doi: 10.1109/ICICT50816.2021.9358774.

[26] S. P. Dewi, A. L. Prasasti, and B. Irawan, "The Study of Baby Crying Analysis Using MFCC and LFCC in Different Classification Methods," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, Jul. 2019, pp. 18–23. doi: 10.1109/ICSIGSYS.2019.8811070.

[27] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech Emotion Recognition with Dual-Sequence LSTM Architecture," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6474–6478. doi: 10.1109/ICASSP40776.2020.9054629.

[28] J. Oruh, S. Viriri, and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," *IEEE Access*, vol. 10, pp. 30069–30079, 2022, doi: 10.1109/ACCESS.2022.3159339.

[29] S. A. Sanjaya and S. Suyoto, "Generating Combination of Biblical Baby Names using Recurrent Neural Network (RNN) and Optimization Comparison," in *2023 4th International Conference on Big Data Analytics and Practices (IBDAP)*, Aug. 2023, pp. 1–5. doi: 10.1109/IBDAP58581.2023.10271934.